Packet Scheduling Algorithm Considering a Minimum Bit Rate for Non-realtime Traffic in an OFDMA/FDD-Based Mobile Internet Access System

Dong-Hoi Kim, Byung-Han Ryu, and Chung-Gu Kang

ABSTRACT—In this letter, we consider a new packet scheduling algorithm for an orthogonal frequency division multiplexing access/frequency division duplex (OFDMA/FDD)-based system, e.g., mobile broadband wireless access or high-speed portable internet systems, in which the radio resources of both time and frequency slots are dynamically shared by all users under a proper scheduling policy. Our design objective is to increase the number of non-realtime service (e.g., WWW) users that can be supported in the system, especially when the minimum bit rate requirement is imposed on them. The simulation results show that our proposed algorithm can provide a significant improvement in the average outage probability performance for the NRT service, i.e., significantly increasing the number of NRT users without much compromising of the cell throughput.

Keywords—Packet scheduling, minimum bit rate, orthogonal frequency division multiplexing access (OFDMA) system, non-realtime traffic, adaptive modulation and coding (AMC).

I. Introduction

The medium access control protocol in an orthogonal frequency division multiple access/frequency division duplex (OFDMA/FDD)-based wireless access system exploits both time-division and frequency-division multiplexing for efficient resource utilization. In particular, a packet scheduler in a medium access control layer will cooperate with adaptive modulation and coding (AMC) which dynamically determines a proper level of modulation enforced with a channel coding scheme of variable rate [1].

In this letter, we consider a new packet scheduling algorithm for an OFDMA/FDD system. Our design objective is to increase the number of non-realtime (NRT) service users that can be supported at the same time in the system, especially when the minimum bit rate (MBR) requirement is imposed on them. As opposed to the existing scheduling algorithms, e.g., the proportional fairness (PF) algorithm (which deals with the non-realtime services as a best effort service class), the proposed algorithm takes the short-term delay constraint into account to meet the design objective.

This letter is organized as follows. Section II describes the system model including a scheduler structure and system parameters. Section III presents the details of the proposed algorithm. Section IV provides the simulation model and results for performance analysis. Concluding remarks and further studies are given in section V.

II. System Model

1. Scheduler Structure

The structure of the packet scheduling system in a typical base station (BS) is composed of three basic blocks: a buffer and QoS management block to store the NRT IP traffic received from the upper layer and manage the corresponding QoS information, the per-sub channel bit rate calculator block to compute the per-subchannel bit rate (PBR) of packets stored in the buffer and QoS management block for each user, and a packet scheduler to execute the actual algorithm. The proposed packet scheduling algorithm is carried out, depending on the per-subchannel bit rate calculated for each user by the per-subchannel bit rate calculator and the status report given by the user equipment (UE). The details of the packet scheduling algorithm is presented in section III.

Manuscript received Mar. 31, 2003; revised Dec. 8, 2003.

Dong-Hoi Kim (phone: +82 42 860 1133, email: donghk@etri.re.kr) and Byung-Han Ryu (email: rubh@etri.re.kr) are with Mobile Telecommunication Research Division, ETRI, Daejeon, Korea.

Chung-Gu Kang (email: ccgkang@korea.ac.kr) is with the College of Information & Communications, Korea University, Seoul, Korea.

2. System Parameters

We consider an OFDMA/FDD system with 1,536 sub-carriers derived from a total bandwidth of 20 MHz. All sub-carriers are shared among all the users in terms of subchannels, which are defined as a subset of the sub-carriers. We have a total of 12 subchannels in the current system, i.e., 128 sub-carriers in each subchannel. Furthermore, we simply assume that all sub-carriers are used for data transmission, not taking the control sections and other overhead into account. All 128 sub-carriers in each subchannel are selected almost in a random manner and thus, we assume that the signal to interference ratio (SIR) measured for all the subchannels of each user is the same. Depending on the instantaneous SIR of each subchannel, the modulation and coding scheme is determined by the predefined AMC table. All the system parameters are summarized in Table 1.

Table 1. System parameters.

Parameter	Value
System	OFDMA/FDD
Downlink channel bandwidth	20 MHz
OFDM symbol duration	100 μs
Total number of subcarriers	1,536
Number of subcarrier per subchannel	128
Number of subchannels	12
Frame period	20 ms
Slot period	1 ms

III. Algorithm Description

1. Conventional Algorithm

Although there are a number of packet scheduling algorithms which are currently used, our proposed scheduling algorithm will be compared with the proportional fairness and modified largest weighted delay first algorithms. We note all of these algorithms are based on the priority metric $\mu_i(t)$, and for each subchannel, the user with the highest priority metric is preferentially served at each scheduling instance *t*. These two scheduling algorithms under consideration are differentiated by the different priority metric $\mu_i(t)$. For example, a priority metric for the PF algorithm is given as follows [2]:

$$\mu_i(t) = r_i(t)/\bar{r}_i(t), \tag{1}$$

where $r_i(t)$ is the current data rate which the BS can support and $\bar{r}_i(t)$ is an exponentially smoothed average of the service rate received by user i up to slot t. Note that both $r_i(t)$ and $\bar{r}_i(t)$ are time-varying with adaptive modulation as they depend on the channel condition. The user with the highest $\mu_i(t)$ will receive the first transmission per subchannel at each decision time. Ties are broken randomly. Any user for whom there is no data to send is not taken into scheduling. The exponentially smoothed average of the service rate, $\bar{r}_i(t)$, is updated by $\bar{r}_i(t+1) = (1-1/T_c)\cdot\bar{r}_i(t) + \ 1/T_c \cdot b_i(t)$, where the parameter T_c is the time constant for exponential filtering; e.g., T_c= 1000 slots (1 sec). We use $b_i(t)$ to denote the current transmission rate of user *i* per subchannel at time *t*. A user who is not currently receiving service has $b_i(t) = 0$. Even users for whom the scheduler has no data to send also get their average rate updated.

The modified largest weighted delay first (M-LWDF) algorithm takes the maximum delay requirement for each user *i*, $W_{\max}^i$, into account [3]. For every subchannel at slot time *t*, the corresponding priority metric is given as follows:

$$\mu_i(t) = a_i W_i(t)\cdot R_i(t)/\overline{R}_i, \tag{2}$$

where $W_i(t)$ is the head-of-the-line packet delay for queue *i*, $R_i(t)$ is the channel capacity with respect to user *i*, and $\overline{R}_i$ is the average channel rate with respect to user *i*. Also, $a_i = -\log\delta_i / W_{\max}^i$ with δ_i being the maximum probability of $W_i(t)$ exceeding $W_{\max}^i$. The user with the highest $\mu_i(t)$ will receive the first transmission per subchannel at each decision time. If we assume that all users belong to the same service class with the same QoS requirement and thus, $W_{\max}^i$ and δ_i are the same for all users, then (2) can be simplified to

$$\mu_i(t) = W_i(t)\cdot R_i(t)/\overline{R}_i. \tag{3}$$

2. Proposed Algorithm

A. Per-Subchannel Bit Rate

For each user *i* with the total length of $L_i(t)$ bits for the backlogged packets in the queue at time slot *t*, a *per-subchannel* bit rate at slot time *t* is defined as $\widetilde{P}_i(t) = L_i(t)/\widetilde{W}_i(t)$, where $\widetilde{W}_i(t)$ is the waiting time for a head-of-line packet for user *i* in each subchannel. Conceptually, the per-subchannel bit rate is an effective bit rate that is virtually warranted if each user *i* per subchannel is served in the current slot time *t*, i.e., taking $\widetilde{W}_i(t)$ seconds to serve $L_i(t)$ -bit packets, resulting in an effective rate of $\widetilde{P}_i(t) = L_i(t)/\widetilde{W}_i(t)$ bps. An important performance measure to take into account is the relatively short waiting time with respect to the length of the backlogged packets per subchannel. It plays an essential role in reflecting the instantaneous bandwidth requirement of the NRT service, subject to the required minimum data rate, into the priority metric.

B. Packet Scheduling Algorithm Considering MBR

The average bit rate of a packet call is defined as [4]:

$$\overline{A} = A/\tau, \tag{4}$$

where τ is the duration of the packet call and A is the amount of bits actually serviced to user *i* during a packet call. Each of the NRT service users is imposed with a required minimum data rate, $R^i_{\min}$, as a new QoS parameter. The design objective of our proposed scheduling algorithm is to find a new priority metric that can increase the number of NRT service users satisfying the minimum bit rate requirement. For simplicity, we assume that all data users have the same required MBR, i.e., $R^i_{\min} = R_{\min}, \forall i.$

In order to increase the total number of WWW users that meet their MBR requirements, the sharing of a constant data rate is desired while giving a higher priority to the lower bit-rate user. For every subchannel at each slot *t*, therefore, our new priority metric is now defined as follows:

$$\mu_i(t) = R_i(t) / [\overline{R}_i \cdot \widetilde{P}_i(t)]. \quad (5)$$

As the per-subchannel bit rate tends to indicate the instantaneous resource over-allocation, the number of non-realtime service users satisfying the MBR requirement can be increased by the above priority metric. Note that a user with no backlogged packet to send, i.e., $\widetilde{P}_i(t) = 0$, is not taken into account for scheduling. When the $L_i(t)$ is large and $\widetilde{W}_i(t)$ is short, the per-subchannel bit rate $\widetilde{P}_i(t)$ increases. As long as $\widetilde{P}_i(t)$ is larger, a value of the priority metric per subchannel for user *i* must be decreased accordingly and vice versa. Also, when this rule is performed in accordance with any other rule considering the channel capacity $R_i(t)$ and the average channel rate $\overline{R}_i$, it increases the number of the users satisfying the MBR.

C. Subchannel Allocation for OFDMA

We assume that there are 12 subchannels (a subset of sub-carriers in an OFDMA system) available for each time slot. Assuming that all sub-carriers in each subchannel are randomly distributed in the frequency domain, it is acceptable that the same power is assigned to every subchannel in the average sense. More specifically, we assume that each subchannel is set at 1 W, i.e., a total power of 12 W is allocated for the BS, which is equally distributed over all subchannels. We note that a dynamic bit loading and power allocation scheme can be applicable to a more advanced OFDMA system, but it is beyond our current scope.

For each subchannel of a given slot, the user with the largest value of priority metric (5), say user *k*, is identified and then, the packets remaining in the corresponding queue are served therein. If the capacity of the corresponding subchannel in the current slot is not large enough to fully accommodate the packets remaining in the queue of user *k*, then the remaining packets that are not yet processed must be served later. Note that the total number of bits that can be transmitted by one subchannel in each slot is governed by an AMC scheme, which depends on the SIR reported by the UE in case of downlink scheduling. Once all the subchannels in each slot are exhausted, the same process repeats in the next slot.

IV. Performance Analysis

1. Performance Measure

The physical layer shall be capable of adapting the modulation, coding and power levels to accommodate RF signal deterioration between the BS and UEs. Furthermore, the air interface may use appropriate automatic repeat request schemes to ensure the packet error rate. We assume that the BS has perfect channel knowledge. To ensure the packet error rate performance, we follow the AMC table given in Table 2 [5], which specifies the target SIR for the different modulation and coding rates to ensure the average packet error rate of 1%. In other words, the packet error rate is now the performance measure under various channel conditions, which subsequently determines a proper level of modulation order and channel coding rate.

Table 2. Transmission mode with convolutionally-coded modulation [5].

Target SIR (dB)	Modulation scheme	Coding rate
1.5	BPSK	1/2
4.0	QPSK	1/2
7.0	QPSK	3/4
11.0	16 QAM	9/16
13.5	16 QAM	3/4
18.5	64 QAM	3/4

In the simulation, the SIR is measured as the ratio of signal power to interference for each user. More specifically, the SIR of the *n*-th subchannel allocated for the *m*-th user in a cell *k* can be represented as

$$SIR(m,n) = P_{mnk} \Big/ \sum\nolimits_{j \neq k} P_{mnj}, \quad (6)$$

where P_{mnk} is the signal power received by the *m*-th user for the *n*-th downlink subchannel in the cell *k*. All subchannels are allocated with equal power, i.e., 1 W, while the corresponding received power P_{mnk} is governed by the propagation model. In the current simulation, only the pass loss is taken into account, which it is given in Table 4.

For the NRT packet service, subject to a minimum bit rate requirement, we introduce a new performance measure to quantify an event when the average bit rate serviced within a packet call of NRT users becomes less than the MBR. More specifically, it is given in terms of an average outage probability, defined as follows:

$$P_{out} = \mathrm{Num}\left(\overline{A} < R_{\min}\right) / N_{total}, \quad (7)$$

where N_{total} is the total number of packet calls generated throughout the simulation, and $\mathrm{Num}(\overline{A} < R_{\min})$ is the number of packet calls with $\overline{A}$ less than $R_{\min}$.

In practice, we want to maintain the average outage probability for all users below a certain level, i.e., there will be a threshold average outage probability of all users, which is denoted as η. In fact, our design objective is to maximize the number of NRT users that meet $P_{out} \le \eta$.

2. Simulation Model and Parameters

A. Traffic Model

Figure 1 shows a non-realtime traffic model of world wide web (WWW) services, given in terms of their sessions, which are defined by the time intervals of the WWW browsing usage. Each session consists of a number of packet calls, each of which also comprises a burst of packets. In other words, a burst of packets makes up one packet call, which is a major characteristic in the NRT traffic source.

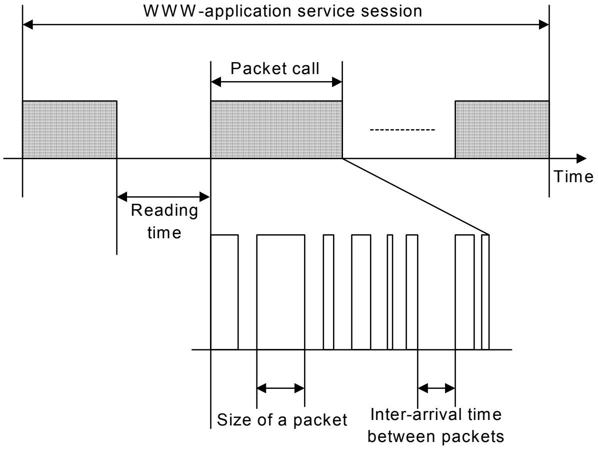
WWW-application service session
Packet call
Time
Reading time
Size of a packet
Inter-arrival time between packets

Fig. 1. Traffic model for NRT packet service session.

In Fig. 1, the size of a packet can be modeled as the Pareto distribution, while the number of packet calls per session and the number of packets within a packet call follow the geometric distributions with the different means. On the other hand, the inter-arrival time between the packets and the reading time between packet calls follow the geometric distributions. Denoting the packet size by a random variable *p*, a truncated packet size *L* is modeled as a truncated Pareto distribution and given as follows:

$$f_L(x) = \begin{cases} \alpha k^{\alpha} / x^{\alpha+1}, & k \le x \le m \\ \beta, & x = m \end{cases}, \tag{8}$$

where parameter α denotes the shape of the probability density function for packet size, and the minimum and maximum allowed packet sizes are denoted by *k* and *m*. Also, β is the probability that $x > m$, which is given as $\beta = (k/m)^{\alpha}$, $\alpha > 1$. The average packet size can be calculated as

$$\begin{aligned} \mu_L &= \int_{-\infty}^{\infty} x f_L(x) = \int_{k}^{m} x \cdot \alpha k^{\alpha} / x^{\alpha+1}\, dx + m(k/m)^{\alpha} \\ &= [\alpha k - m(k/m)^{\alpha}] / (\alpha - 1). \end{aligned} \tag{9}$$

The specific parameters for the WWW surfing, unconstrained delay data 144 kbps traffic model are given in Table 3 [6].

B. System Model

We consider a hexagonal cell layout with a reference cell and 6 surrounding cells in the first tier, each with an omni-directional antenna. Even if more tiers can be taken into account for accurate assessment of other cell interference, it is not quite essential in our studies, simply because the current analysis is focused on the fair comparison between the proposed and existing schemes. A long radius of each cell is fixed at 1 km. The uniformly-distributed mobile stations are moving with the velocity of uniform distribution in a random direction. For the WWW surfing, unconstrained delay data 144 kbps traffic model, the MBR requirement is set to $R_{\min} = 50$ kbps. The BS transmission power is given by 12 W, which is equally distributed among all 12 subchannels. Simulation parameters for the system model are summarized in Table 4.

Only the payload of each packet is considered, e.g., excluding the additional bits for the header. To overload the system for a short duration, the reading time between packet calls is set to 2.6592 seconds without conforming to the parameter given in Table 3. All results in Figs. 2 and 3 are obtained for simulation durations of 100 second-long actual sessions.

3. Results and Discussion

Figure 2 shows the average outage probability as the number of WWW users increases. It is obvious that the proposed scheme outperforms in outage performance over the proportionally fairness and modified largest weighted delay first schemes, especially when the number of WWW users increases. For the threshold average outage probability $\eta = 0.1$, the number of WWW users supported by the proposed scheme is increased by almost 20% as compared with those supported by existing schemes.

The total throughput per cell is defined as an effective data rate for the data bits correctly received by all users in each cell. We note from Fig. 3 that total throughput obtained by the proposed scheme is not much different from the other schemes, which implies that our capacity improvement (illustrated by outage performance) is achieved without compromising the throughput performance. It is attributed to the fact that the instantaneous bandwidth requirement

Table 3. WWW-application traffic model parameters.

Information types	Distribution	Distribution parameters
Number of packet calls per session	Geometric	5 packets
Reading time between packets calls	Geometric	412 seconds
Number of packets within a packet call	Geometric	25 packets
Inter-arrival time between packets (within a packet call)	Geometric	0.0277 seconds
Packet size	Truncated Pareto	k = 81.5 bytes, m = 66666 bytes, α=1.1, μ_L = 480 bytes

Table 4. Simulation parameters for system model.

Parameter	Value
User distribution	Uniform
Number of cell	7 cells
Beam pattern	Omni-directional
Cell radius	1 km
User velocity	3 ~ 100 km/s (uniformly distributed)
Cell layout	Hexagonal grid with one tier
Path loss model [7]	L=128.1+37.6log$_{10}R$; R in km
Base station Tx power	12 W

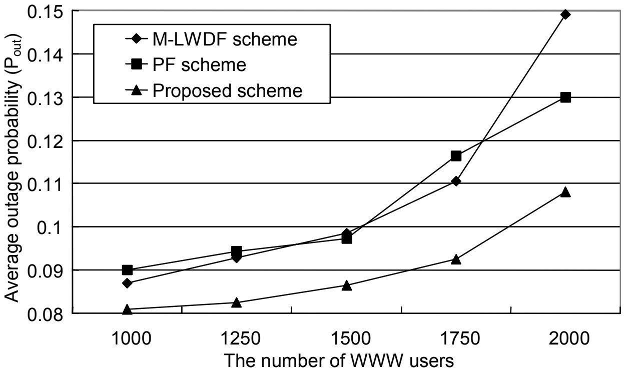

Fig. 2. Average outage performance.

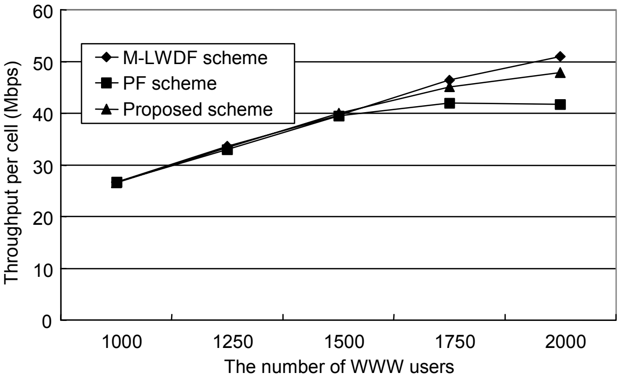

Fig. 3. Per-cell throughput performance.

of the NRT service, subject to the required minimum data rate, is properly reflected into the priority metric.

V. Conclusion

In order to maximize the number of non-realtime service users with the minimum bit rate requirement, we have integrated the per-subchannel bit rate into the priority metric of the scheduling algorithm. As opposed to most of the existing algorithms targeted at either the best-effort service (without any QoS requirements), or the delay-constrained services, our proposed algorithm turns out to be a useful means of providing the required minimum data rate for some services. It has been designed to maximize the number of NRT users subject to a MBR while maintaining the overall throughput as much as the existing algorithms. Simulation studies show that the proposed scheme outperforms other existing schemes in outage performance, conforming to our design objective.

Future work includes performance analysis under mixed traffic scenarios, especially with individual algorithms optimized for different types of service traffic.

References

[1] D. Yoon, D.I. Chang, N.S. Kim, and H.S. Woo, "Linear Diversity Analysis for M-ary Square Quadrature Amplitude Modulation over Nakagami Fading Channels," *ETRI J.*, vol. 25, no. 4, Aug. 2003, pp. 231-237.

[2] A. Jalali, R. Padovani, and R. Pankaj, "Data Throughput of CDMA-HDR a High Efficiency-High Data Rate Personal Comm. Wireless System," *VTC 2000-Spring*, vol. 3, 2000, pp. 1854 -1858.

[3] Matthew Andrews, Krishnan Kumaran, Kavita Ramanan, Alexander Stolyar, and Phil Whiting, "Providing Quality of Service over a Shared Wireless Link," *IEEE Comm. Mag.*, vol. 39, Feb. 2001, pp. 150-154.

[4] 3GPP, *Feasibility Study for OFDM for UTRAN Enhancement (Release 6)*, 3G TR25.892 V0.2.0, Mar. 2001.

[5] Qingwen Liu, Shengli Zhou, and Georgios B. Giannakis, "Cross-Layer Combining of Adaptive Modulation and Coding with Truncated ARQ over Wireless Links," *IEEE Trans. Wireless Comm.*, 2004, Available at http://spincom.ece.umn.edu/papers 04/tw02-534final.pdf.

[6] J.P. Castro, *The UMTS Network and Radio Access Technology*, John Wiley & Sons, Inc., New York, NY, 2001.

[7] 3GPP, *Physical Layer Aspects of UTRA High Speed Downlink Packet Access (Release 2000)*, 3G TR25.848 V4.0.0, Mar. 2001.